Twinning Project

Strengthening the Capacity of the Georgian Statistical System

Component 4: "*Strengthening Social Statistics*"

Sub-component 4.1: "Labour Force Methodologies and Indicators"

MISSION REPORT

Activity: 4.1.B (RS) "Presentation and recommendations on imputation methods and methodological documentation"

Mission carried out by
Aksel Thomsen, Statistics Denmark
Peter Tibert Stoltze, Statistics Denmark

December 3rd to December 16th 2020

Version: Final

Expert contact information

Mr. Aksel Thomsen
Statistics Denmark
Sejrøgade 11
DK-2100 Copenhagen
Denmark
Tel: +45 3917 3593
Email: akt@dst.dk

Mr. Peter Tibert Stoltze
Statistics Denmark
Sejrøgade 11
DK-2100 Copenhagen
Denmark
Tel: +45 3917 3827
Email: psl@dst.dk

Table of contents

1. General comments

This mission report was prepared within the EU Twinning Project "Strengthening the Capacity of Georgian Statistical System". This was the third mission within the sub-component 4.1: "Labour Force Methodologies and Indicators". Due to COVID-19 the mission was carried out as a Remote Session with 3 meetings in the period from December 3 to December 16, 2020. The mission was mainly devoted to presentation and recommendations on imputation methods and methodological documentation in relation to the Georgian Labour Force Survey.

The purposes of the mission were:

- Assessment of the current situation of imputation of LFS data
- Presentation and recommendation to imputation methods in LFS
- Presentation and recommendation to imputation tools in LFS
- Implementation and test of imputation methods
- Discussions and recommendations to the compilation of methodological documentation

The consultants would like to express their gratitude to the Geostat staff who participated in the mission, for the kind support and valuable information received during the mission.

The views and observations stated in this report are those of the consultants and do not necessarily correspond to the views of the European Union, Geostat, Statistics Denmark, or other statistical institutions involved in the implementation of the project.

2. Assessment and results

The mission was conducted by virtue of three remote sessions aptly spaced in time to allow for reflection on the previous meeting and preparation for the next. As a companion to this mission report, three presentations are supplied. They constitute the major part of the contribution from the expert, and thus document the results achieved. In the following we present a very brief account of the presentations (one from each remote session):

Presentation 1: Dealing with non-response and Introduction to imputation
The presentation introduces the concept of non-response, both item non-response and partial non-response. A simple terminology is proposed (the R matrix) which defines response rate on both question level and subject level. Then non-response is considered in a sampling perspective, and why raising design weights with the inverse of the response propensity is not always a sufficient way to deal with non-response. Imputation is introduced as a possible measure against the detrimental effect of non-response, and various classes of imputation methods are introduced. Finally, the added uncertainty following from using partially imputed data is highlighted.

Presentation 2: Donor based imputation methods with application in R
The presentation starts with a recap on the general approach to donor-based imputation. Two distinct approaches, nearest neighbour and (stratified) random draw, are discussed: KNN (k nearest neighbors) and PMM (predictive mean matching) in the most detail. The application of these methods are demonstrated using the R software and a simulated (or synthetic) data set having structure somewhat equal to a LFS. At first it is shown how easy it is to make indicators for response / non-response using the `tidyverse` approach in R. Defining a response matrix is a prerequisite for analysing the response pattern and coming up with a suitable imputation strategy. The demonstrated imputation strategy is a two-step procedure applying both deductive (rule based) imputation and donor-based imputation (both predictive mean matching and random hot-deck). The strategy is implemented using the `simputation` framework.

Presentation 3: Georgian LFS and donor imputation

This presentation is based on actual data from the Georgian LFS as provided to the experts in anonymized form. The presentation is in the form of a R notebook which consists of both text, code, and results. The actual document that produces the notebook was already distributed to the Georgian colleagues to be further explored.

The notebook describes a simple albeit complete workflow with the following phases:

1. Read and prepare data
2. Define non-response in the data set
3. Analyse non-response patterns
4. Choose imputation method (for a single variable)
5. Perform imputation of missing values in the survey (for the selected variable)

Regarding phase 4, the evaluation of competing imputation methods is based on a standard approach from statistical learning whereby the observed data is split into a training data set and a test data set. More elaborate methods involving n-fold cross-validation is briefly mentioned.

3. Conclusions and follow up

- The Georgian Labour Force Survey is very well organized with suitable IT tools for managing data collection. As of now, the rate of non-response is very small, and it is probably safe to ignore it for most variables.
- Unfortunately, the rate of non-response is expected to increase in the future due to new data collection methodology, so even if the discussed imputation methods strictly speaking do not make a big difference now, they can become very important in the near future.
- Even if the discussed methods are not tied to a specific software, implementation from scratch can be a very time-consuming endeavor, and probably it will be more efficient to spent time on learning how to use ready-made implementations in existing software packages, like the `simputation` package in the R software.

Actions needed for moving forward:

Action	Deadline	Responsible person
Further training in using the R software, by trying to apply imputation methods to the LFS using R	February 2021	Geostat

Annex 1. Terms of Reference

EU Twinning Project GE 16 ENI ST 06 18

October 5th – October 13th 2020

Component 4: Strengthening Social Statistics

Sub-Component 4.1: Labour Force Methodologies and Indicators

Mandatory results and benchmarks for sub-component 4.1
- Improved Labour Force methodologies and indicators

Indicators of Achievement (baseline and targets):
- Availability of current LFS questionnaire
  - **Baseline**: 2019 – LFS methodology (Questionnaire) is outdated
  - **Target**: End of 2020 – Methodology updated and changes ready for implementation in collection instrument according to ILO methodology

- Availability of systematic approach to missing data and non-response at data collection stage
  - **Baseline**: 2019 – No systematic approach to missing data
  - **Target**: November 2020 – Imputation methodologies (handling of missing data) introduced

- Number of staff with sufficient knowledge on imputation methods
  - **Baseline**: 2019 – 0
  - **Target**: November 2020 – At least 3 staff members trained

Activity 4.1.B (RS): Presentation and recommendations on imputation methods and methodological documentation

1. Purpose of the activity
- Assessment of the current situation of imputation of LFS data
- Presentation and recommendation to imputation methods in LFS
- Presentation and recommendation to imputation tools in LFS
- Implementation and test of imputation methods
- Discussions and recommendations to the compilation of methodological documentation

2. Expected output of the activity
- Current situation of imputation methods in LFS assessed
- Imputation methods in LFS presented
- Recommendation on imputation methods and tools
- Imputation methods and tools are partly tested
- The frame and the content of methodological documentation has been discussed
- Mission report written

Annex 2. Persons met

Geostat
Mr. Vasil Tsakadze, Head of Social Statistics Department
Mrs. Irma Gvilava, Head of Labour Statistics Division
Mr. Zezva Sanikidze, Senior Specialist of Labour Statistics Division
Mr. Tsotne Balakhashvili, Senior Specialist of Labour Statistics Division
Mr. Irakli Guguchia, Senior Specialist of Labour Statistics Division
Ms. Lili Chedia, Senior Specialist of Labour Statistics Division
Mrs. Salome Esadze, Senior Specialist of Labour
Mr. Elizbar Sadradze, Senior Specialist of Labour

RTA Twinning Team
Mr. Steen Bielefeldt Pedersen, Resident Twinning Advisor
Ms. Nino Grdzelishvili, Resident Twinning Advisor Translator